R Virtual laboratory (RvLab) - Statistical Analysis Functions for biodiversity studies

DATA ANALYSIS

Anastasis Oulas (oulas@hcmr.gr)

Christos Arvanitidis (arvanitidis@hcmr.gr)

Overview

RvLab has concentrated its focus on the majority of the functions supported by the vegan CRAN package, which provides methodologies for the analysis of ecological communities. It has tools for analyzing ecological diversity, and for the multivariate analysis of communities (NMDS, pCCA, pRDA etc.), such as diversity analysis, species abundance modelling, analysis of species richness, ordination, support functions for ordination (dissimilarity indices, extended dissimilarities, Procrustes analysis, ordination diagnostics, permutation tests), dissimilarity analyses (ANOVA using dissimilarities, ANOSIM, MRPP, BIOENV, Mantel and partial Mantel tests) and others. RvLab has also extended its support to the optimization of primitive operations that are commonly used by the aforementioned vegan functions.

Requirements analysis and profiling

An extensive requirements analysis and profiling have been carried out on the functions of the vegan package in order to determine their demands in terms of computational effort, memory usage during execution (primary or secondary), and encoding methodology. R's profiling tools, such as the *profr*, *proftools, grid,* and *Rgrapviz* packages, have supported this analysis.

Tools for optimisation of execution

The analysis conducted in the previous step revealed ample possibilities for improving the performance of certain computationally intensive vegan functions. It also pointed out the need to exploit alternative means of storage as, in certain cases, the bi-products during computation require memory resources that few computers can support.

With respect to parallelisation, RvLab developers investigated the features of popular packages, such as *snow, multicore,* and *parallel.* Towards this goal, packages that provide interfaces to *MPI for R* (*Rmpi*) were utilised. Examples of these include *Rmpi* and the *pbdR* packages.

Regarding the handling of memory, an external database—PostgreSQL—was utilised. R packages, such as *dplyr* and *RPostgreSQL* were used to connect with the database in order to store and receive the necessary data.

RvLab interface

R's environment offers a convenient way to construct the RvLab interface through an online web application that provides a user-friendly graphical interface to improve the efficiency and execution of RvLab functions. The online application was implemented by combining a series of web development languages such as HTML and PHP, thus creating a web interface that is directly linked to the PC cluster at HCMR. The alpha version of RvLab is available at http://rvlab.portal.lifewatchgreece.eu/. RvLab is available to users after logging into the system. On the main page, users can find links to a comprehensive tutorial on how to operate the basic functions of RvLab and how to navigate the web application.

Expected advantages

1. **Big data manipulation** (overcome memory barriers);
2. **Improved computational speed** through task segmentation, multi-cores, and a cluster computing environment at HCMR, recently upgraded from LifeWatch); and
3. **Develop an efficient and friendly user interface** for analysis of ecological community data.

Applicability

R scripts for all functions as well as the HTML, PHP and JavaScript codes, are available as supporting information via GitHub, and can be provided upon request. Functions require formatted data occurrence or abundance data and a set of environmental variables or related information. Test datasets are available through the RvLab interface, alongside a reference manual describing the applicability of the virtual laboratory.

Potential users

Currently, the RvLab is available for the scientific community, researchers, academics and students, as well as other users. RvLab is available both for computers, mobiles, and tablets. In the near future, we would like to connect with several research providers, such as environmental data providers, and would like to serve LifeWatch infrastructure and related infrastructure associated with EU BON.

Case studies

Case Study #1

We have applied the biodiversity summary plotting function to 35 soft lagoon macrobenthic samples, which were collected in the context of the pan-European Marine Biodiversity Observation System (COST ACTION). Abundance data are available for thousands of species; the plots below show an overview of some of the visual information available via RvLab (**Figures 1a and 1b**).

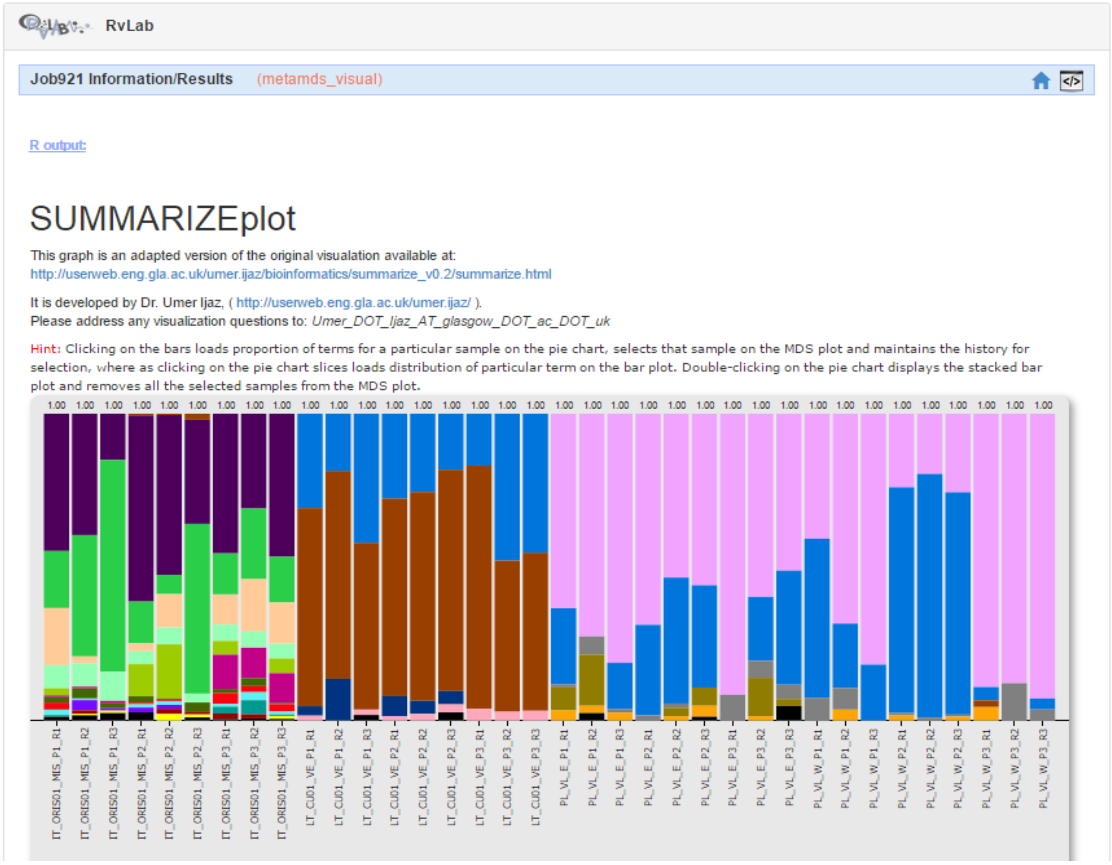

Figure 1(a). Biodiversity Summary Function RvLab result: bar chart showing taxonomic diversity by sample and colour coded by most abundant macrobenthic species.

.

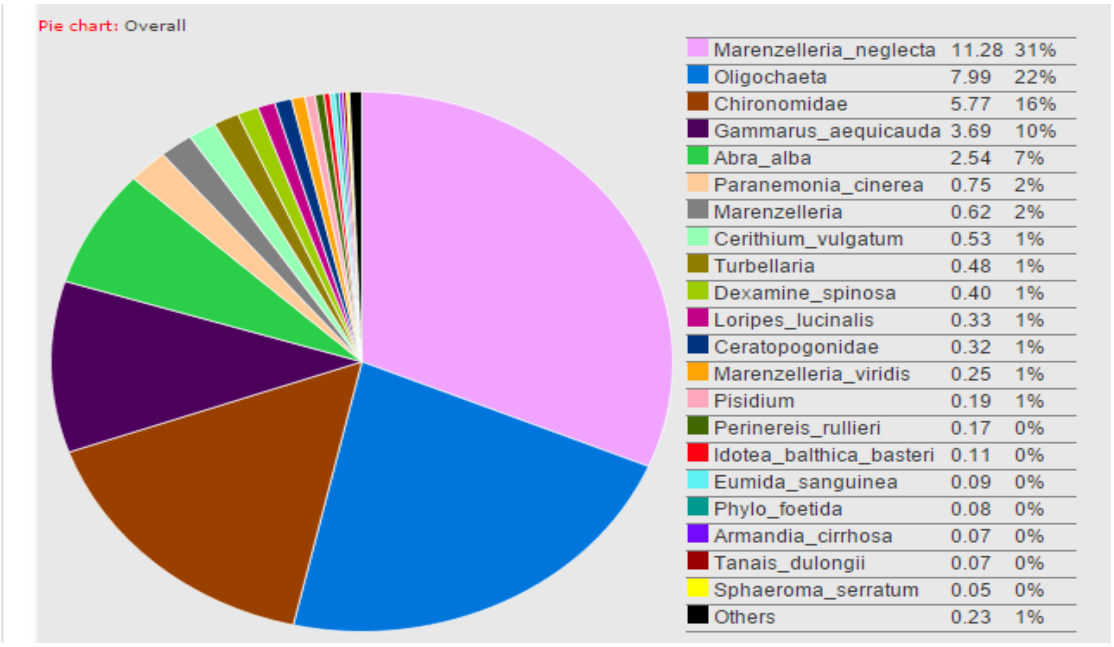

Figure 1(b). Biodiversity Summary Function RvLab result: pie chart showing an overall biodiversity distribution for all samples analysed.

Case Study #2

The taxonomic indices function has been applied to an abundance dataset of more than 150 microbial community samples, collected from multiple locations all over the globe. The biodiversity mapping function shows the geographical location of all samples according to user specified coordinates. Further results are portrayed in an interactive manner. Results are colour coded by the abundance of Operational Taxonomic Units (OTUs) as calculated via the Shannon Index (**Figure 2**).

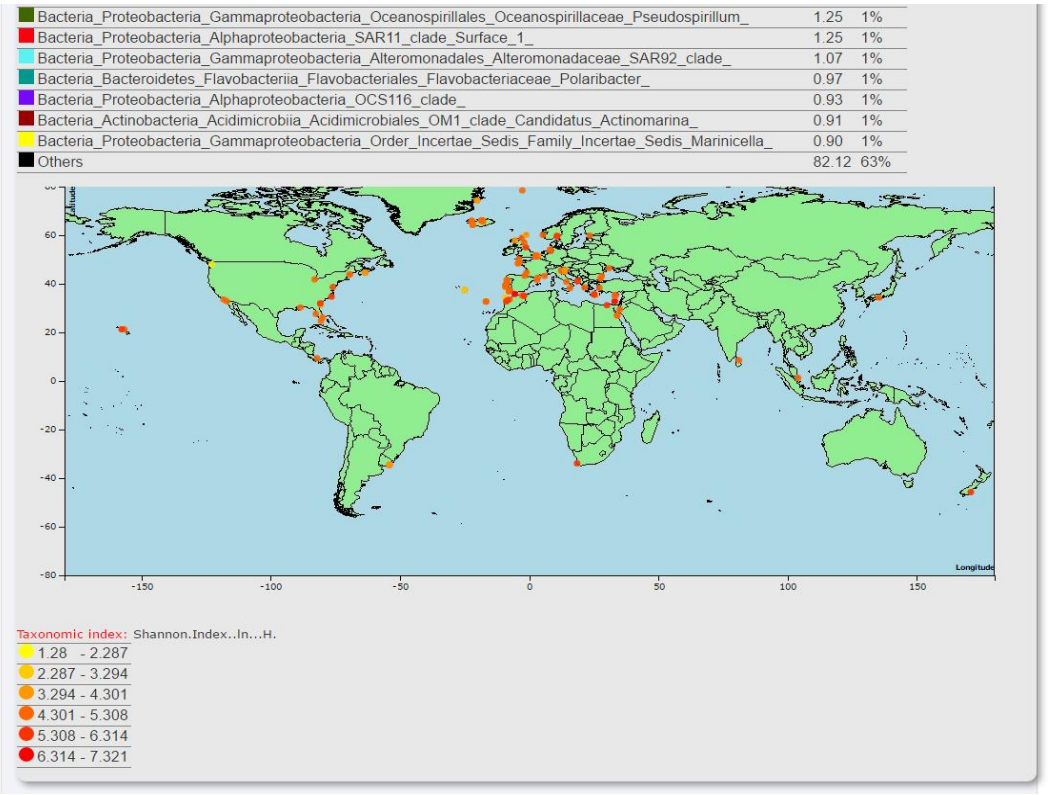

Figure 2. RvLab results of the Microbial OTU Diversity Mapping Function: Shannon's Index values calculated from ~150 samples from all over the world are mapped and colour coded.

Case Study #3

The heatcloud function has been applied to the same microbial OTU abundance dataset of more than 150 samples, analysed in Case Study #2. The function allows for a quick analysis of species' abundances through the development of a heatmap and corresponding tag clouds whereby the most abundance OTUs are colour coded and greater abundances portrayed by increasing font size (**Figure 3**).

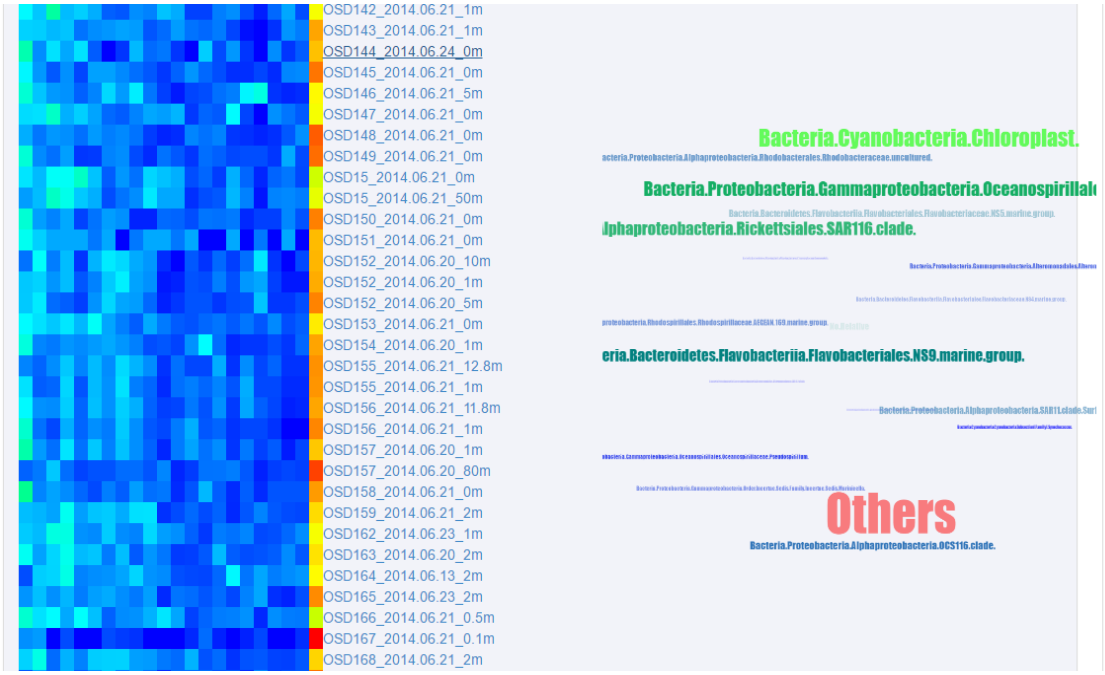

Figure 3. Heatcloud results from RvLab. Microbial OTU abundances from ~150 samples are visualised as a heatmap and associated tag clouds.